

Towards Semantic Libraries

The Container, the Content and the Contenders

Prof. Dr. Stefan Gradmann
Humboldt-Universität zu Berlin / School of Library and Information Science
stefan.gradmann@ibi.hu-berlin.de

... à la carte: A Three-Course Menu



- **Hors d'oeuvre:** Where are we coming from?
Basics of Library Functionality
- **Main course:** Where are we going to?
Semantic (Digital) Libraries
- **Dessert:** Who else is going there?
Partners and Contenders



A faded, light blue background image of a large, multi-story building with many windows, likely a library or university building.

Hors d'Oeuvre

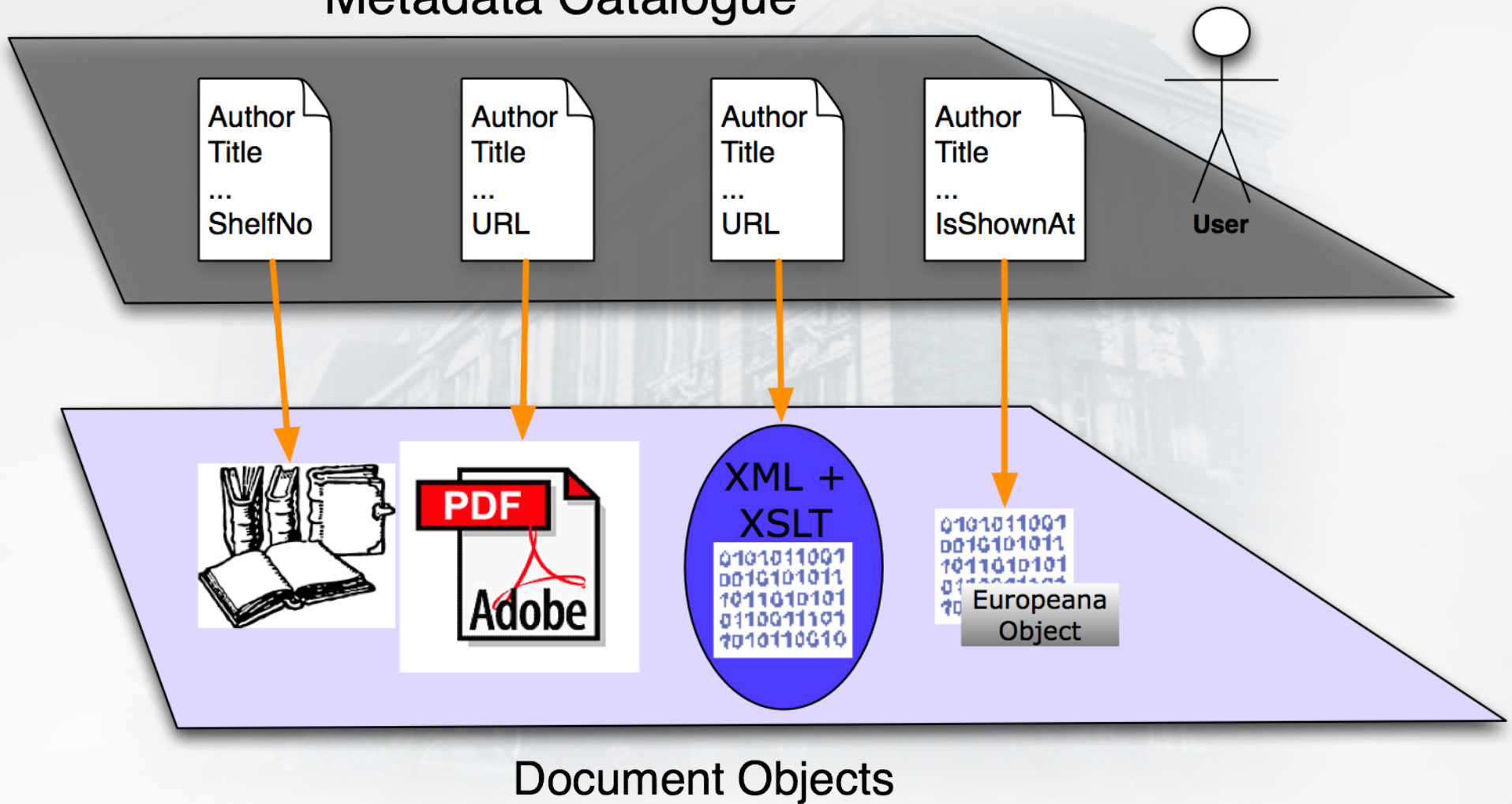
Where are we coming from?

Basics of Library Functionality

Library Functional Principles (1)



Metadata Catalogue



Library Functional Principles (2)



- **Mediating access** to information objects via **catalogues**
- **Mediating links** as pointers from metadata to objects
- Objects are part of a library **collection**
 - An object to be used within a library typically is part of this library's collection
- Internal processing logic: focus on
 - objects as information **containers**,
 - not so much on the **content** of these containers
- **Ingestion, storage, description** and **retrieval** of information objects as functional macro-primitives



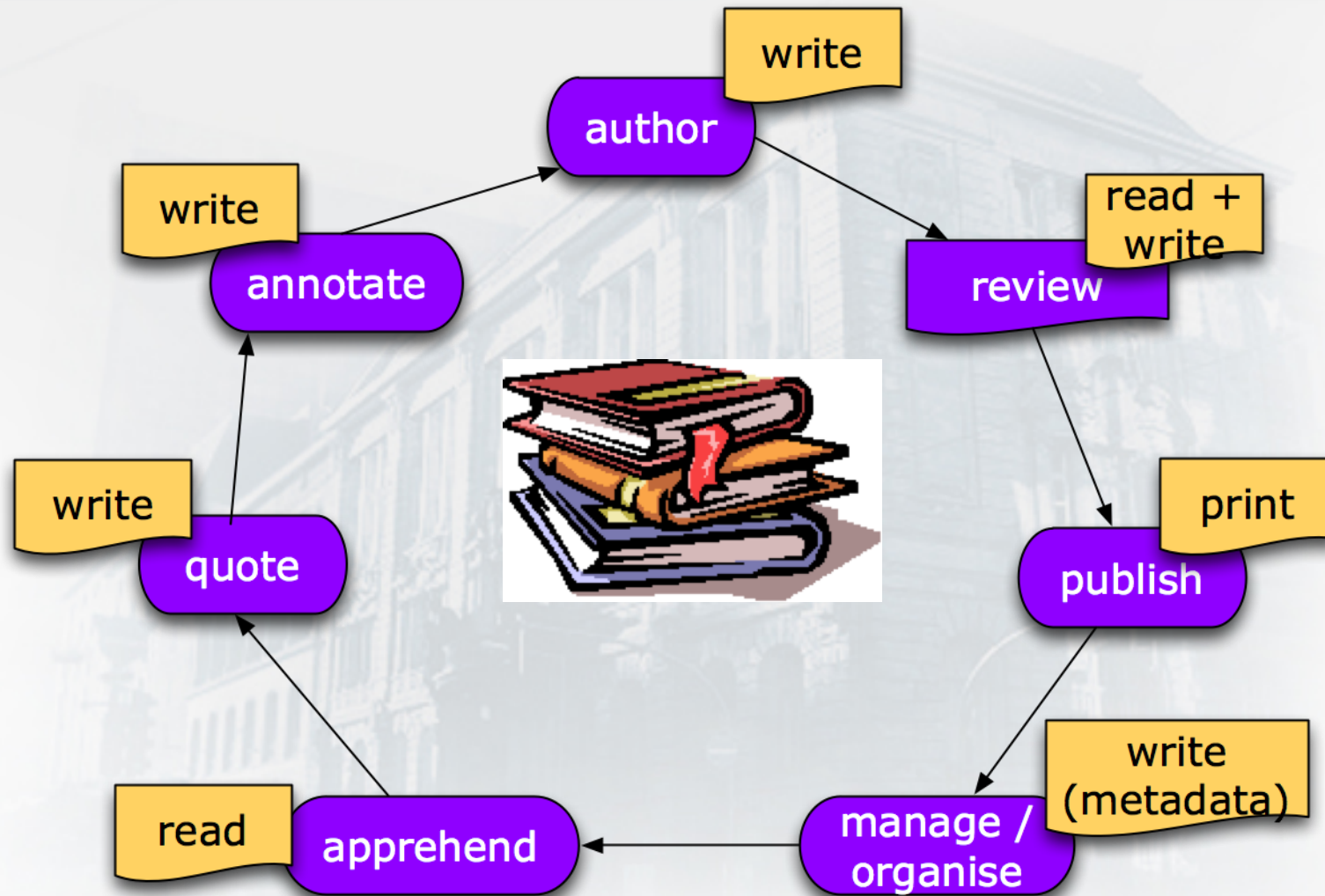
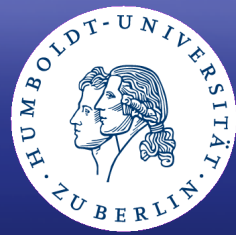
A faded, light blue background image of a large, multi-story building with many windows, likely a university building.

Main Course

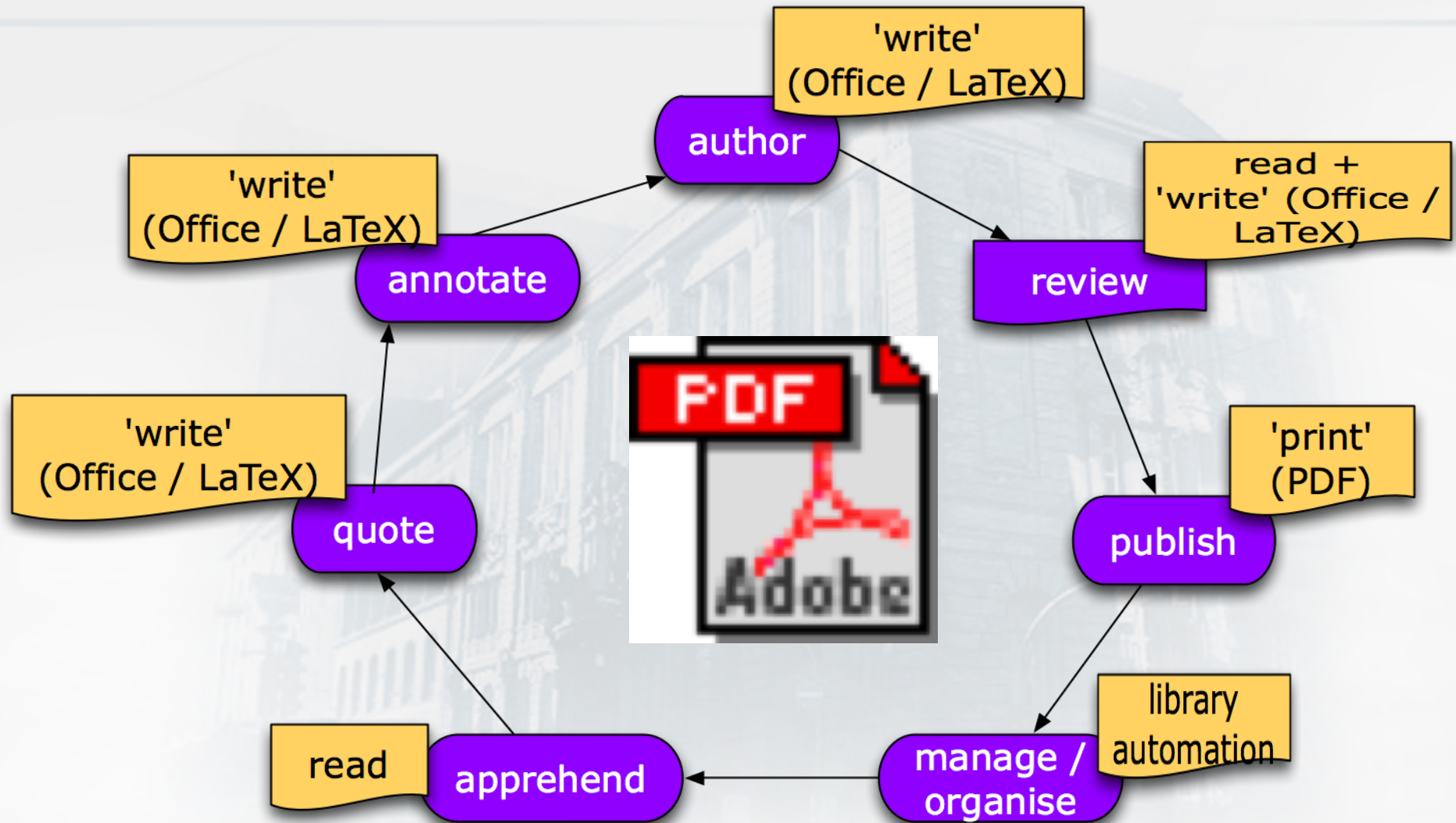
Where are we going to?

Towards the Semantic Library

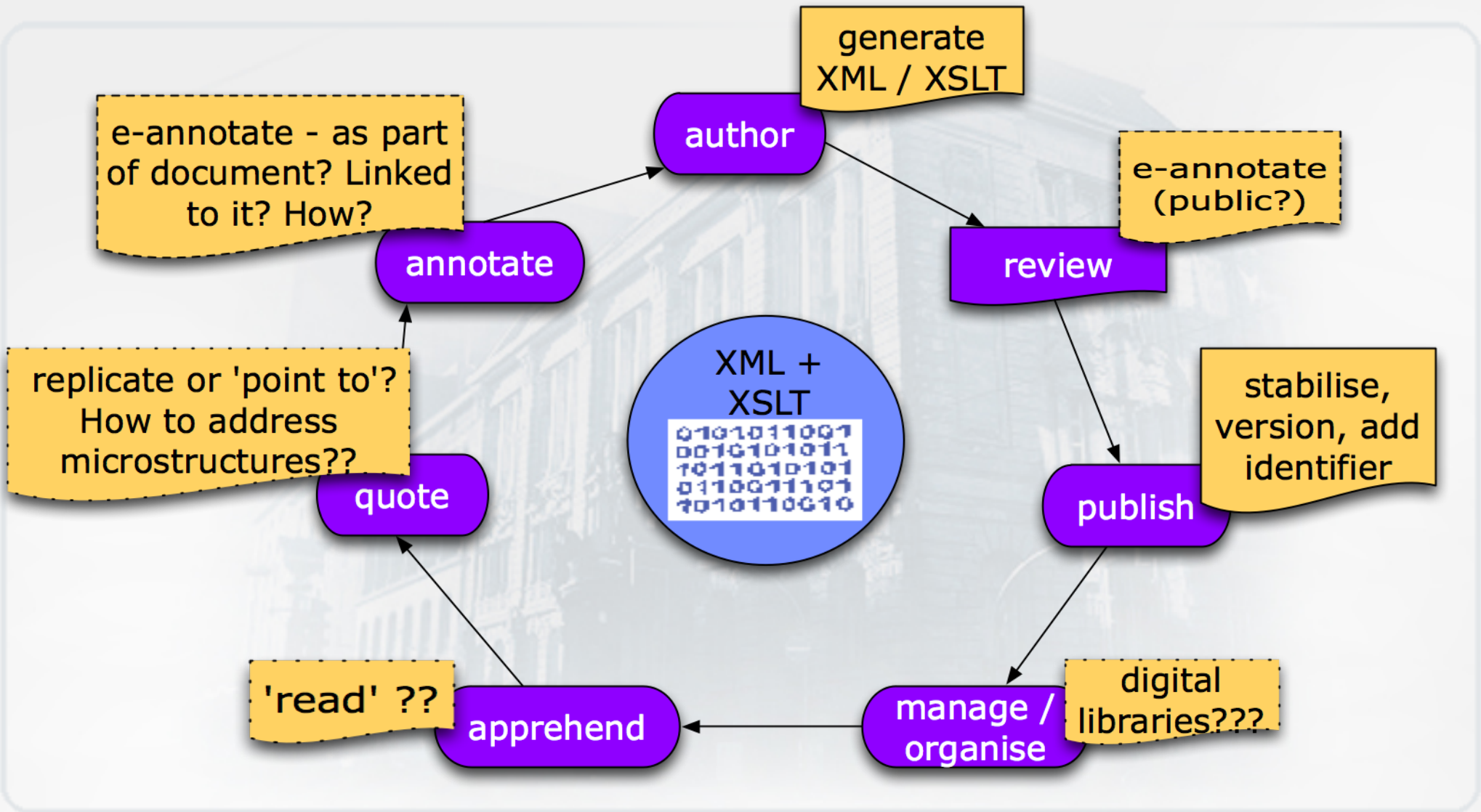
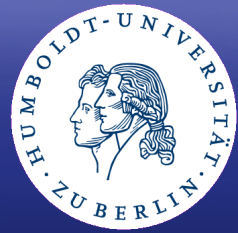
Linear Document Continuum in the Gutenberg galaxy



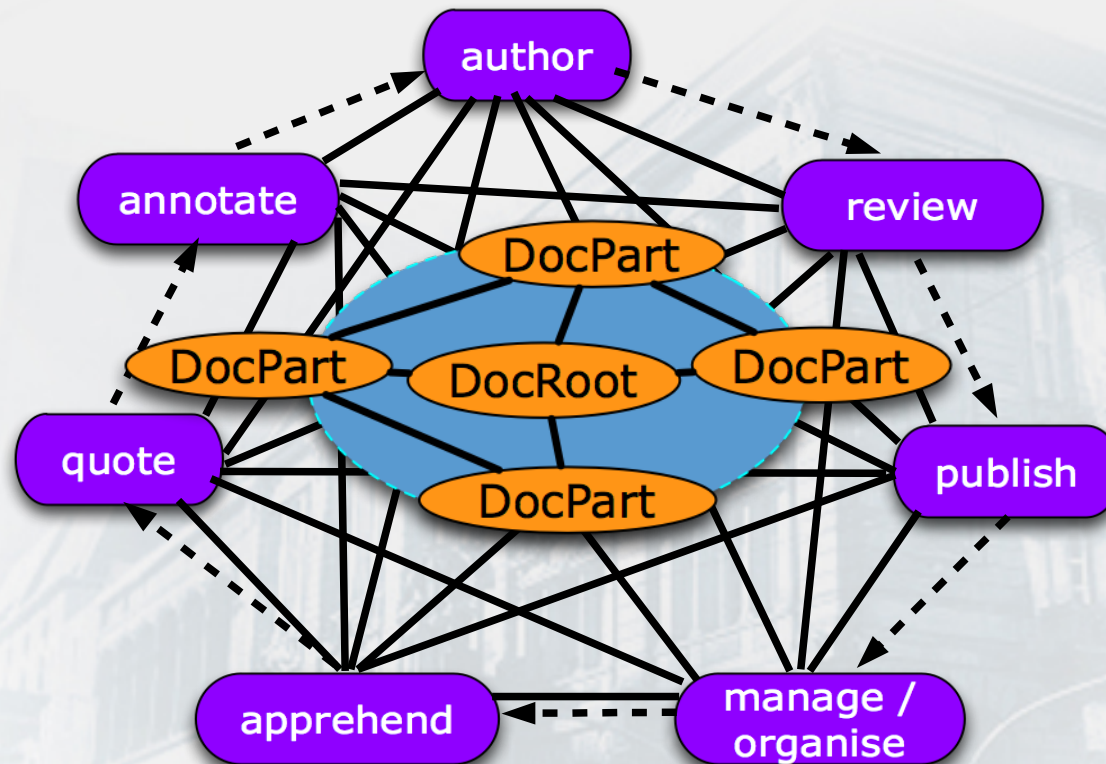
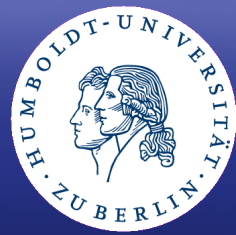
Linear Document Continuum in emulation mode



Linear Document Continuum going digital (entering Turing galaxy)

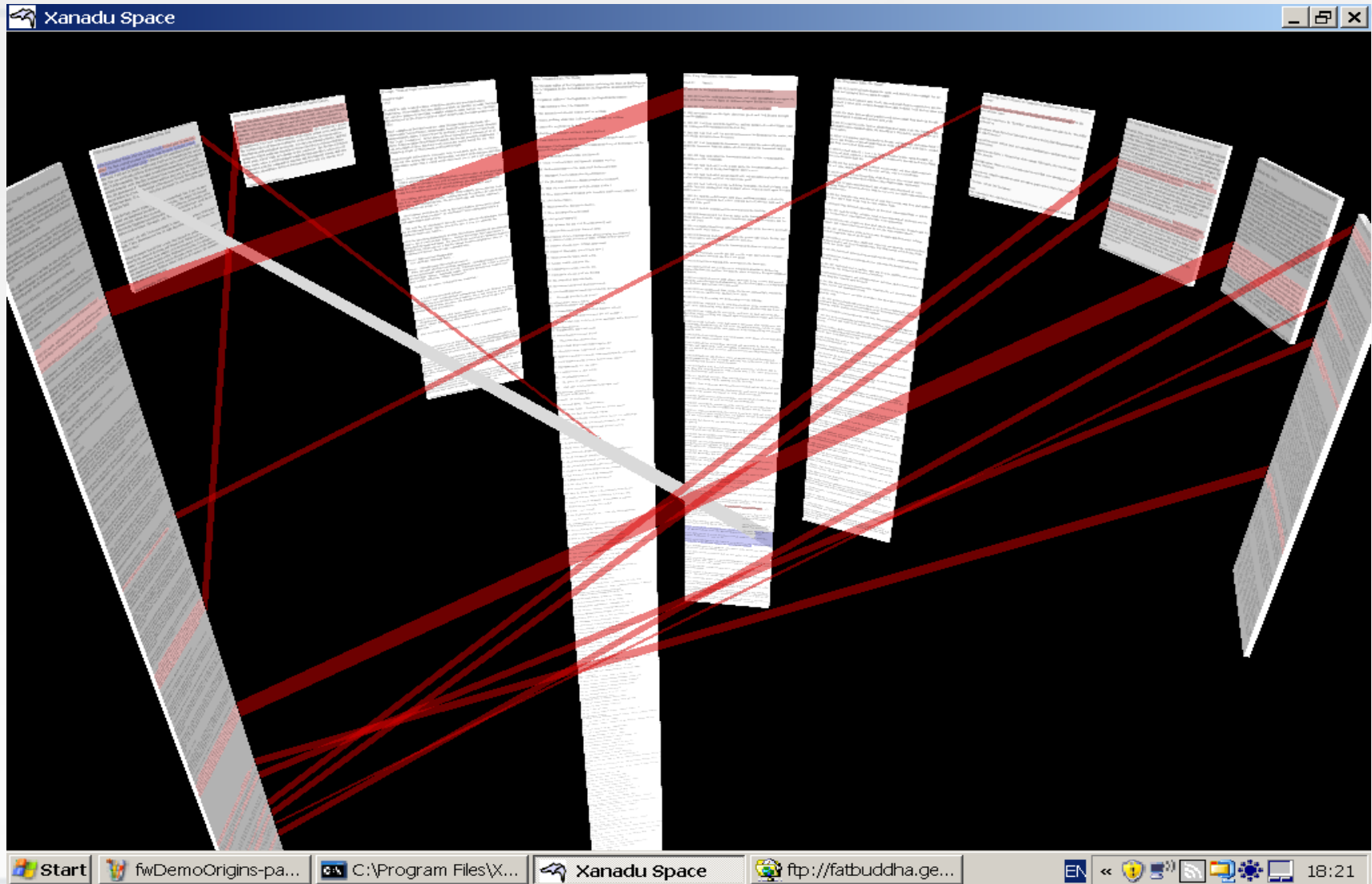


Web Based Scholarly Working Continuum a triple paradigm shift: Beyond Documents

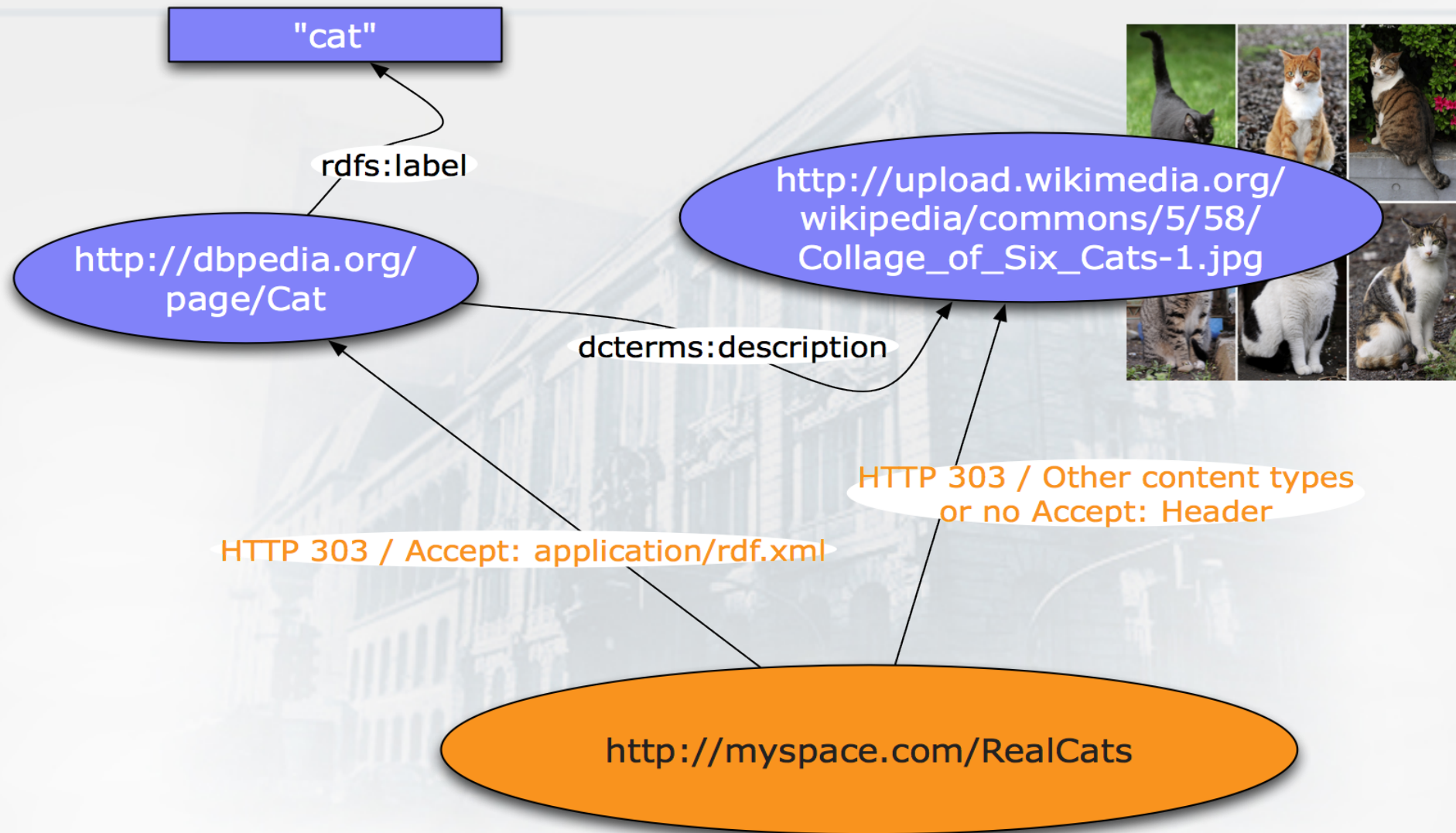
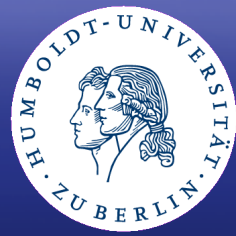


- Decreasing functional determination by traditional cultural techniques
- Disintegration of the linear / circular functional paradigma
- Erosion of the monolithic document notion in hypertext paradigms

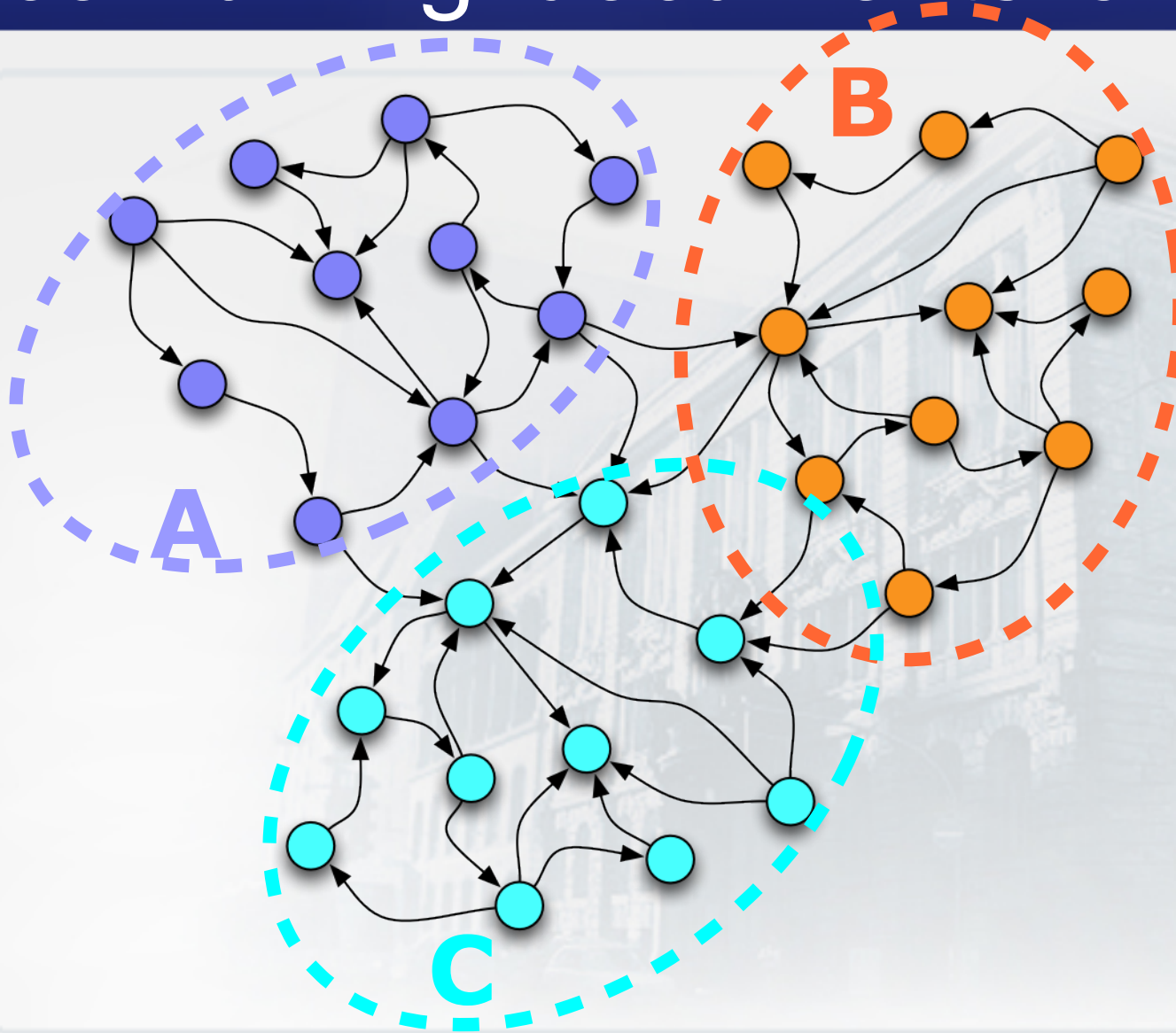
Ted Nelson's Xanadu: radicalised Hypertext ...



... the web of 'documents' extended with a web of 'things' ...



... and 'publication' aggregations combining 'documents' and 'things'



- Where do resource aggregations 'start'? Where do they 'end'?
- And what constitutes document boundaries??
- And which node was connected to which one at a given time???

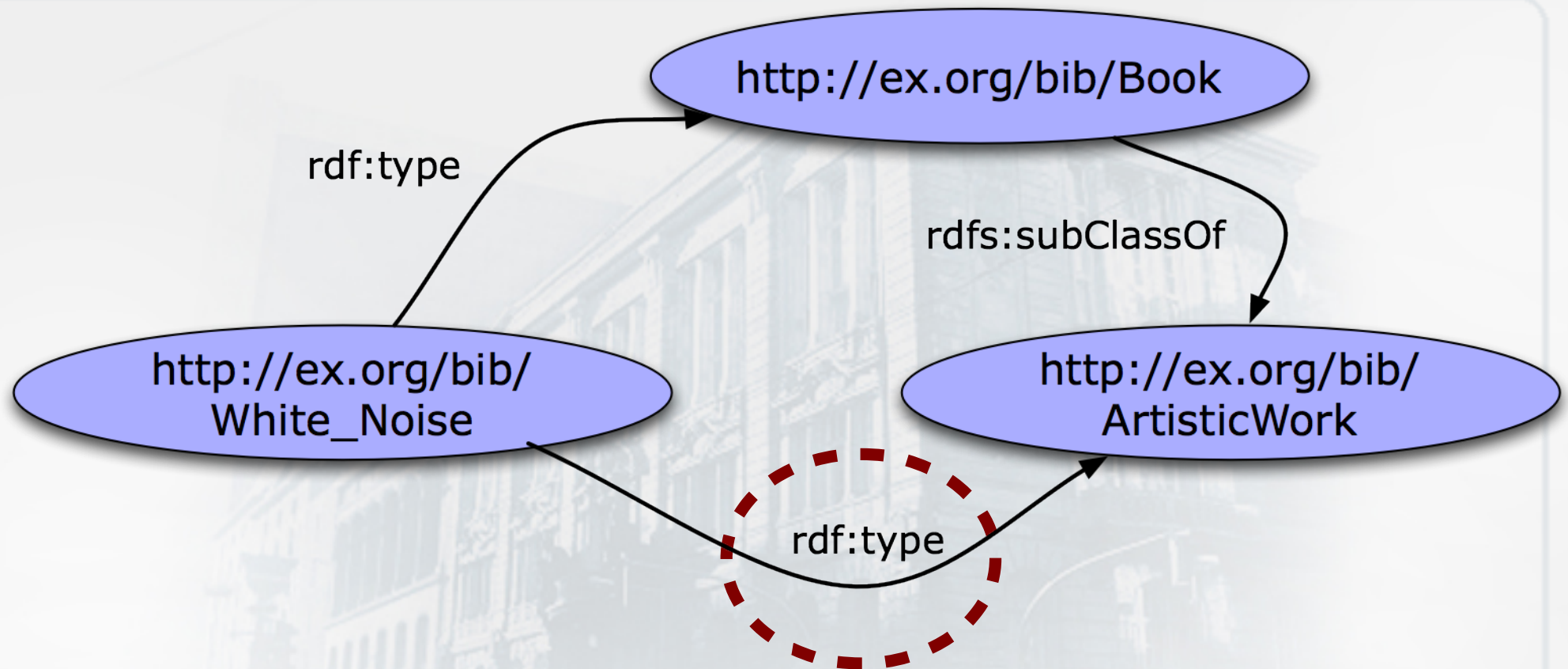
Machines can reason on triple sets!



Some reasoning preconditions ...



... and an automated inference!



There is quite some potential for generating scholarly heuristics here!

The use of Inferences

Citation: van Haagen HHHBM, 't Hoen PAC, Botelho Bovo A, de Morrée A, van Mulligen EM, et al. (2009) Novel Protein-Protein Interactions Inferred from Literature Context. PLoS ONE 4(11): e7894. doi:10.1371/journal.pone.0007894 / Example provided by Jan Velterop



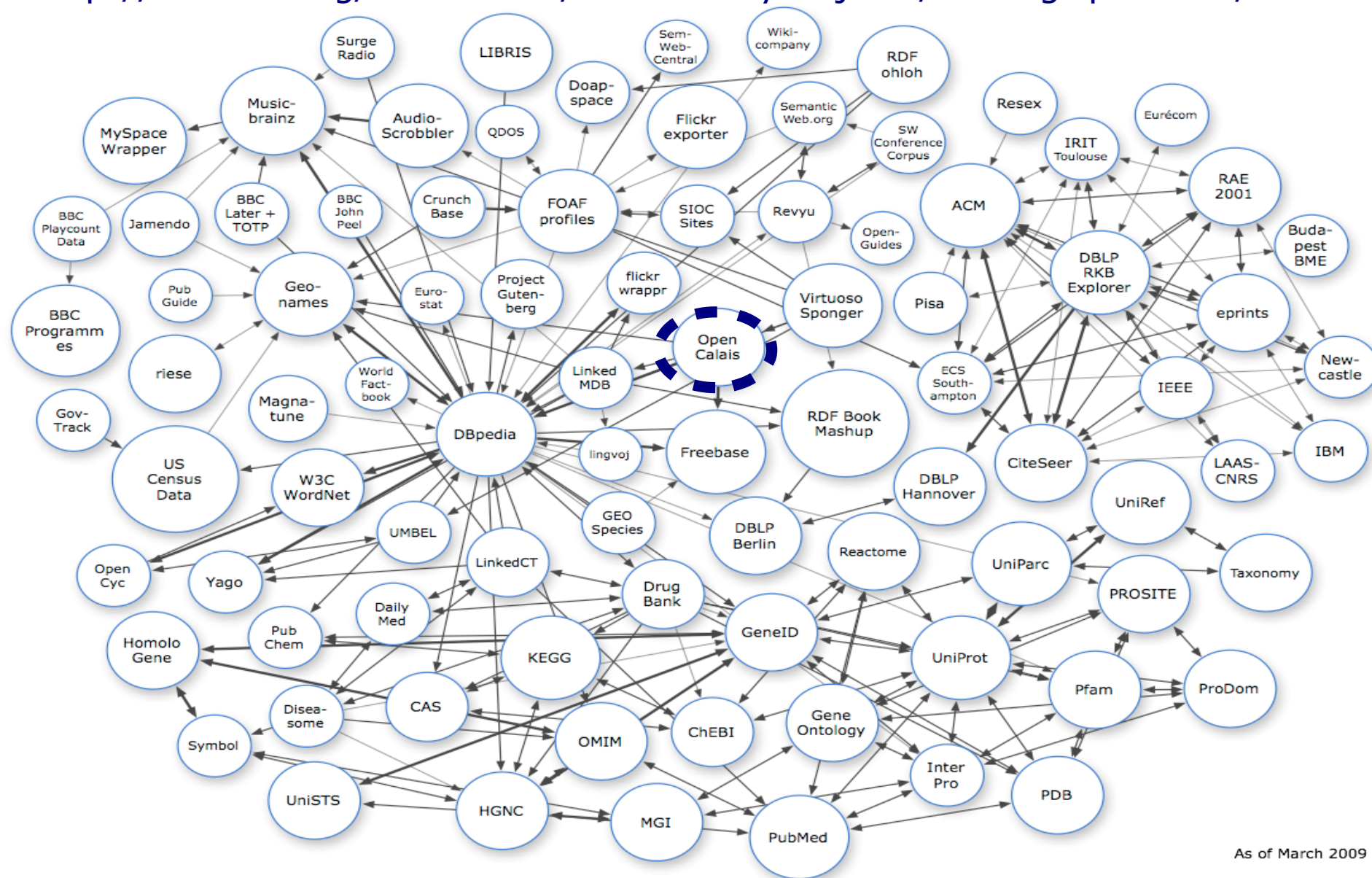
The screenshot shows the PLoS ONE article page. The title 'Novel Protein-Protein Interactions Inferred from Literature Context' is highlighted with a red box. The authors listed are Herman H. H. B. M. van Haagen^{1*}, Peter A. C. 't Hoen¹, Alessandro Botelho Bovo², Antoine de Morrée¹, Erik M. van Mulligen¹, Christine Chichester¹, Jan A. Kors¹, Johan T. den Dunnen¹, Gert-Jan B. van Ommen¹, Silvére M. van der Maarel¹, Vinícius Medina Kern², Barend Mons¹, and Martijn J. Schuemie¹. The abstract states: 'We have developed a method that predicts Protein-Protein Interactions (PPIs) based on the similarity of the context in which proteins appear in literature. This method outperforms previously developed PPI prediction algorithms that rely on the conjunction of two protein names in MEDLINE abstracts. We show'.



LoD: Billions of Triples and Semantic Publishing!



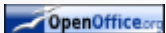
<http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets>



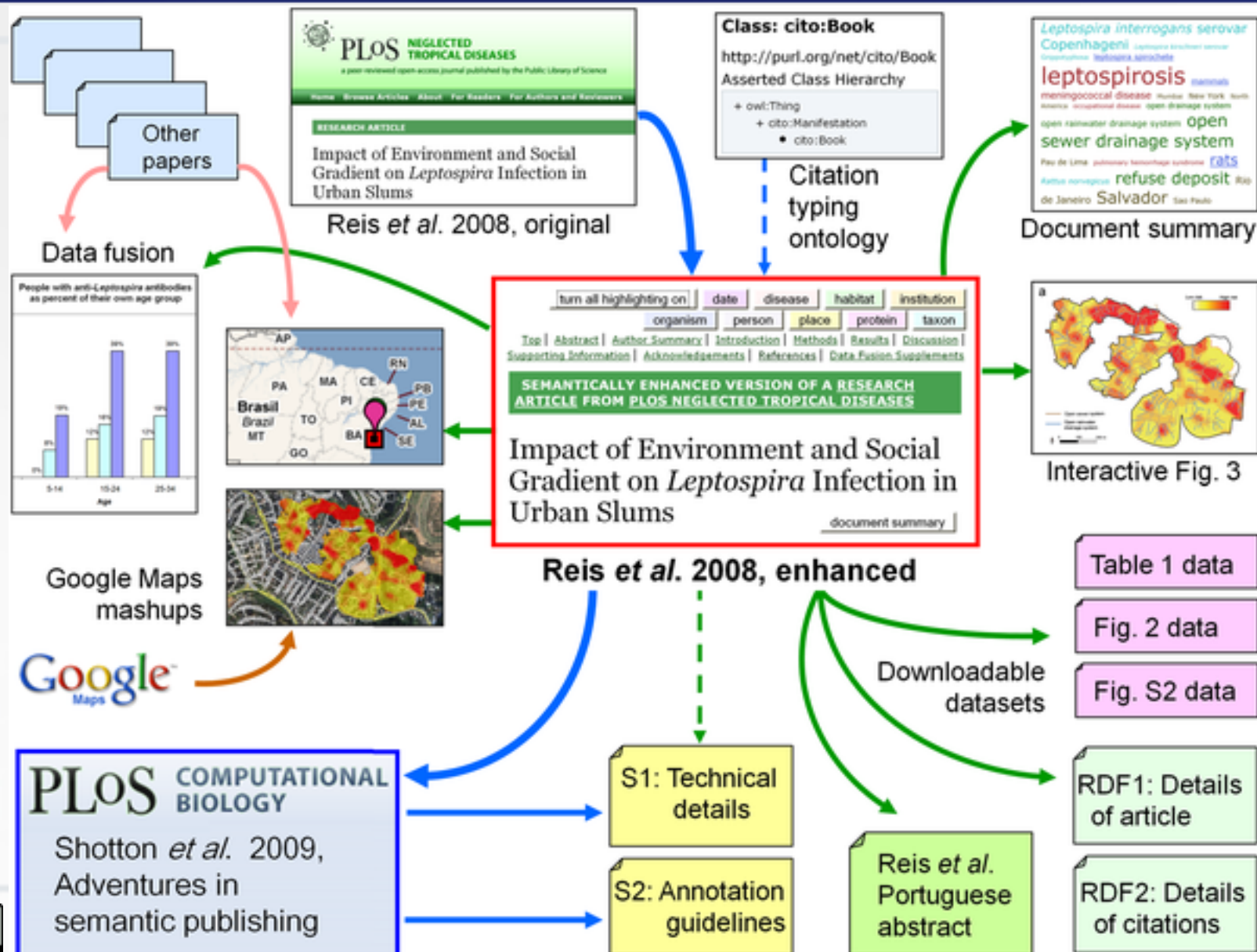
Semantic Publishing as Defined by Shotton



- Shotton et al. (2009b) define semantic publication to include anything that
 - **enhances the meaning** of a published journal article,
 - **facilitates** its automated **discovery**,
 - enables its **linking to semantically related articles**,
 - provides **access to data within the article** in actionable form, or
 - facilitates **integration of data between articles**.
- Example of an enhanced article



Behind the Scean



Semantic Enrichment Tools

- Generic:
 - OpenCalais (<http://www.opencalais.com/>)
 - Temis (<http://www.temis.com/>)
- Specialised:
 - Bio Taxon Finder (http://www.ubio.org/index.php?pagename=xml_services)
 - ConceptWebAlliance (<http://conceptwiki.org>) (Biomedical, Jan Velterop)
- Good critique by Roderic Page:
<http://iphylo.blogspot.com/2009/04/semantic-publishing-towards-real.html>
 - “linking terms to HTML pages doesn't get us much further. Great for humans, not so good for computers.”
 - Too much focus on journal article format!
- → We need a little more!

Books: the Liquid Version

“Turning inked letters into electronic dots that can be read on a screen is simply the first essential step in creating this new library. The **real magic** will come in the second act, as each word in each book is

- cross-linked,
- clustered,
- cited, extracted,
- indexed,
- analyzed,
- annotated,
- remixed,
- reassembled

and woven deeper into the culture than ever before. In the new world of books, every bit informs another; every page reads all the other pages.”

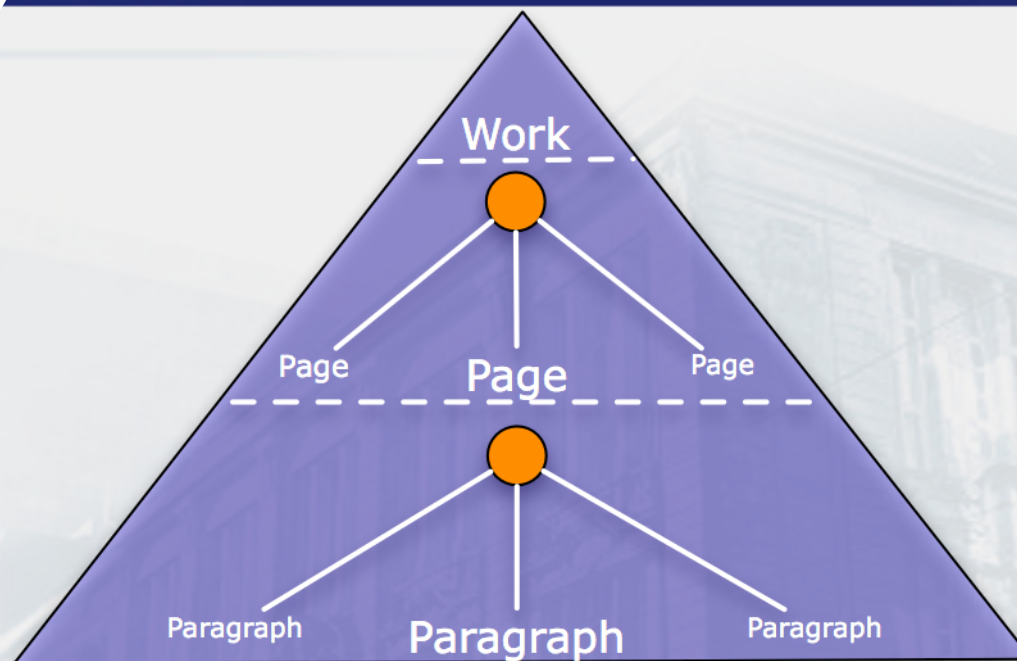
Kevin Kelly, The New York Times Magazine, May 14, 2006

Semantic Micro-Content: PAUX

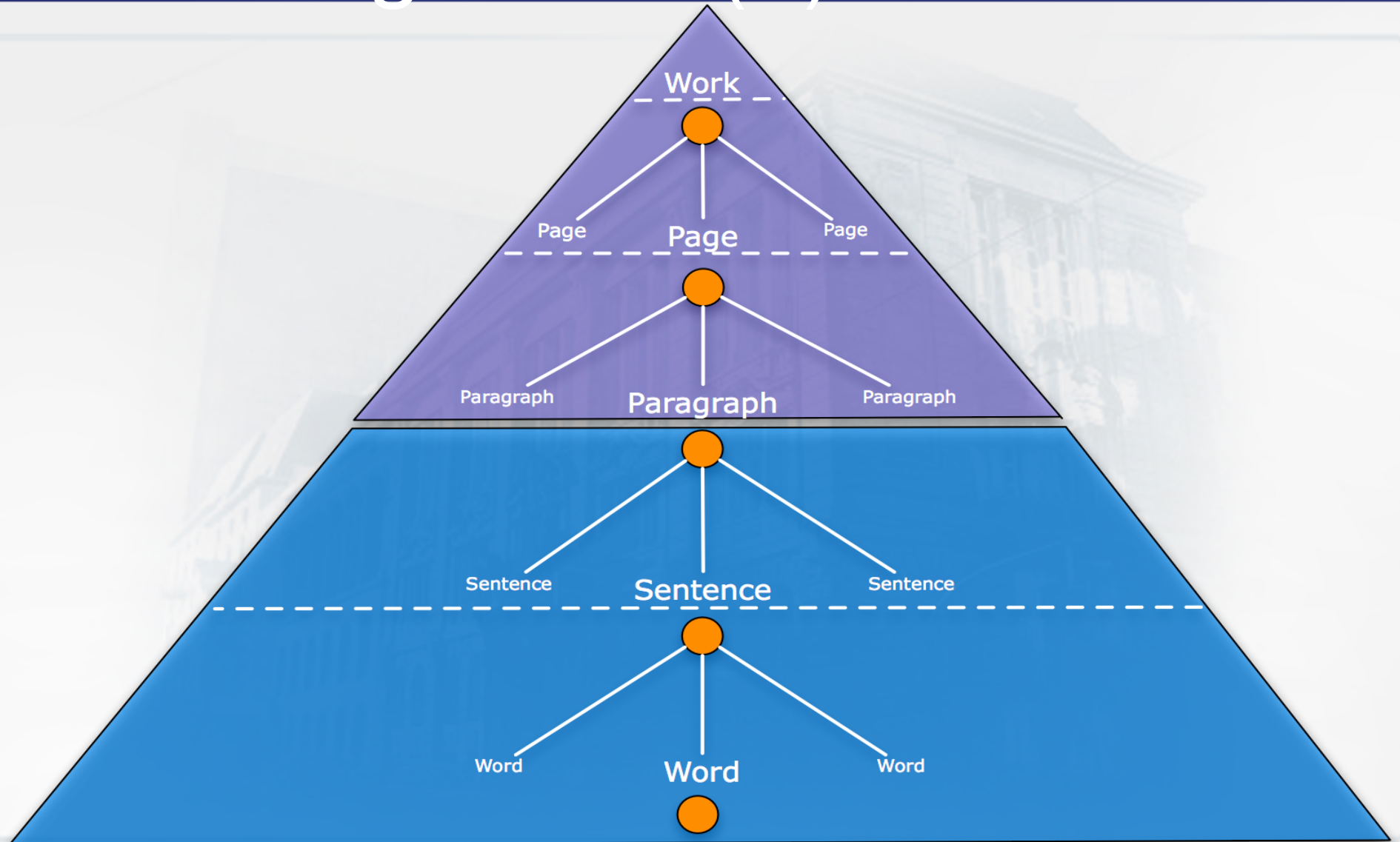


- A **Semantic Wiki**,
- not based on static HTML pages, but instead consisting of **dynamic documents**,
- provided at runtime from **semantic microcontent** (“PAUX-Objects”),
- **semantically linked** by “PAUX-Links”
- Microcontent elements have **HTTP URIs!**
- → PAUX documents can be published as Linked (Open) Data aggregations with maximum granularity: down to **word level**.
- **PAUX creates “liquid books”**
- PAUX (<http://www.paux.de>): origins in eLearning

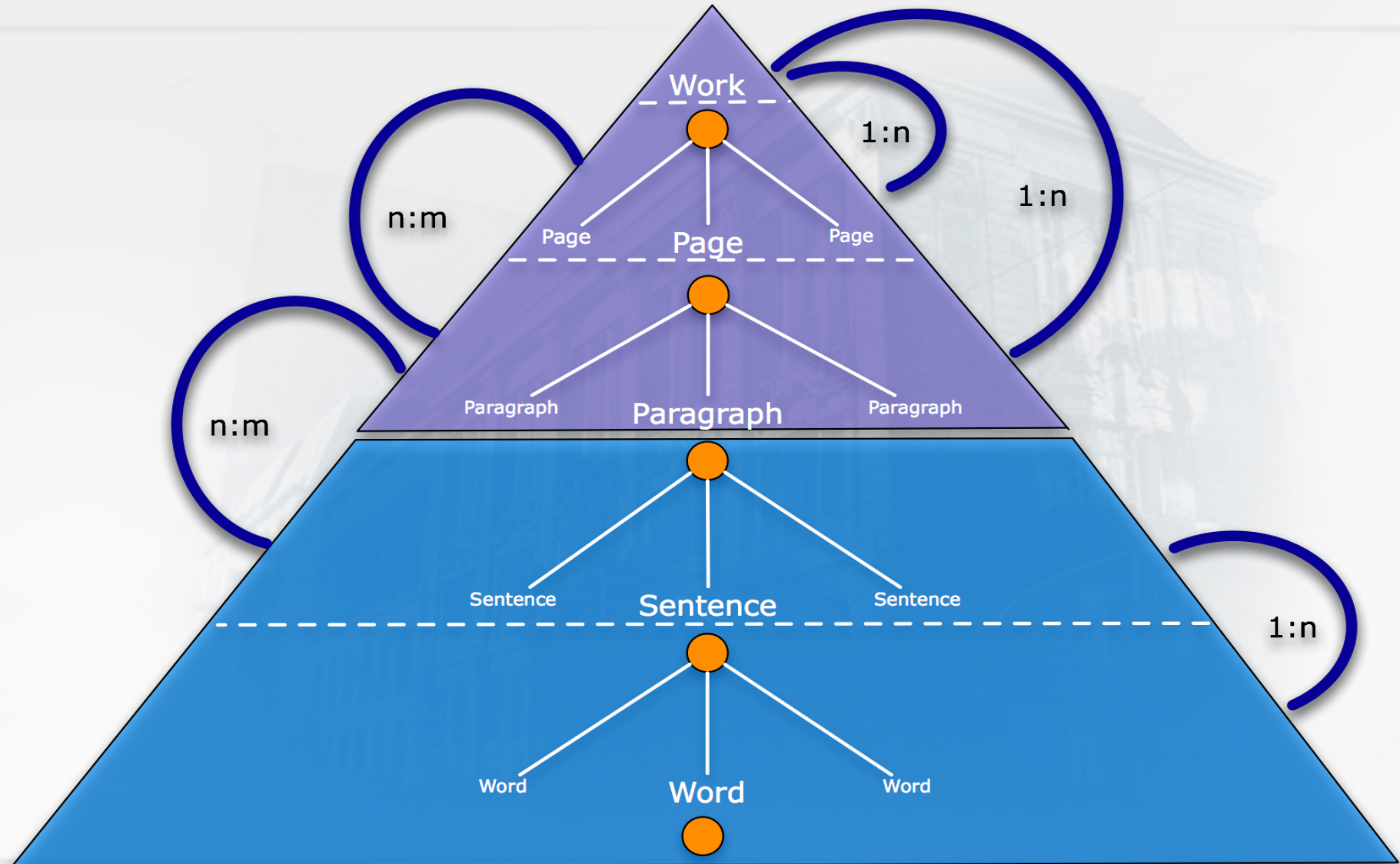
Granular Semantic Publishing: Paux (1)



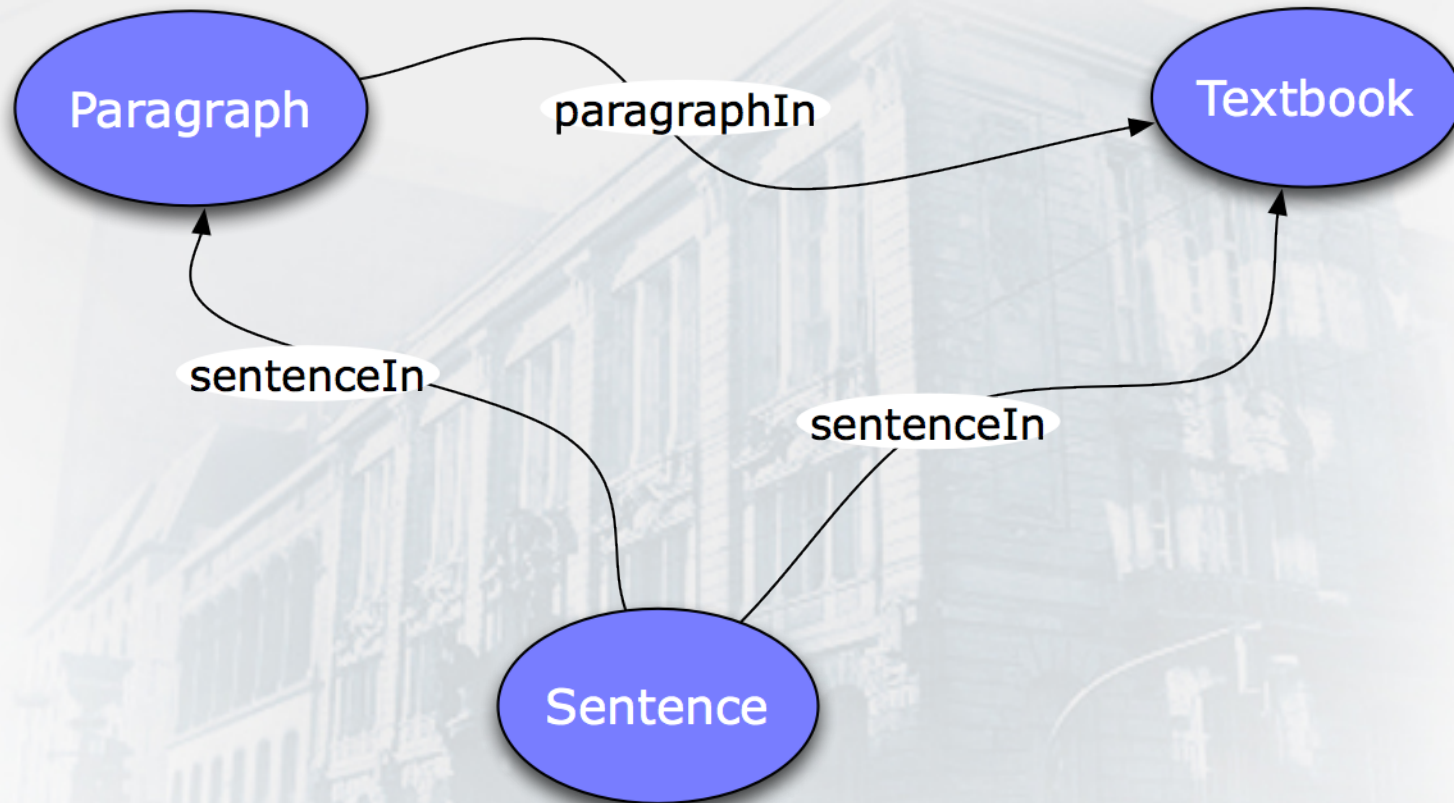
Very Granular Semantic Publishing: Paux (2)



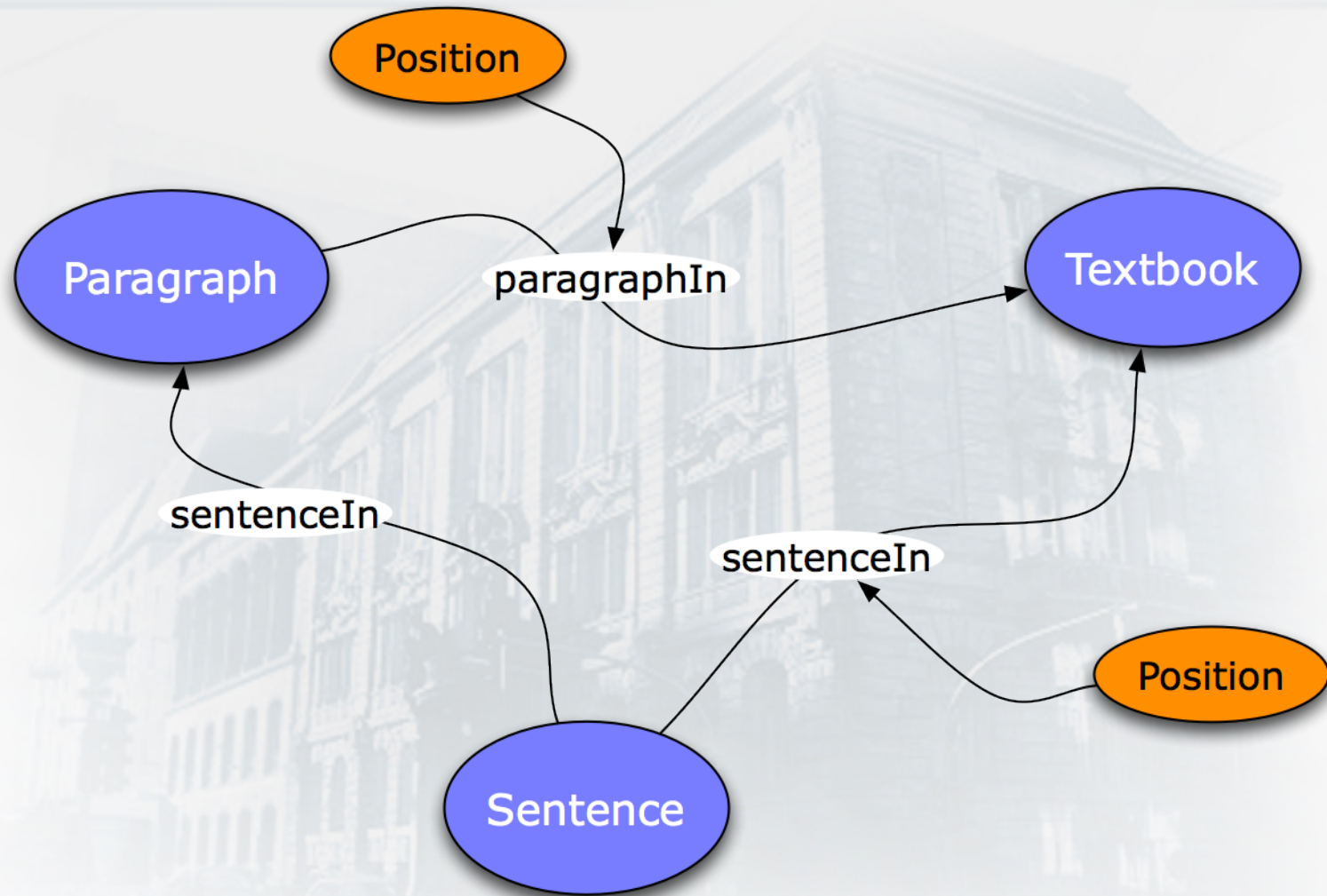
Semantic Publishing: Paux (3)



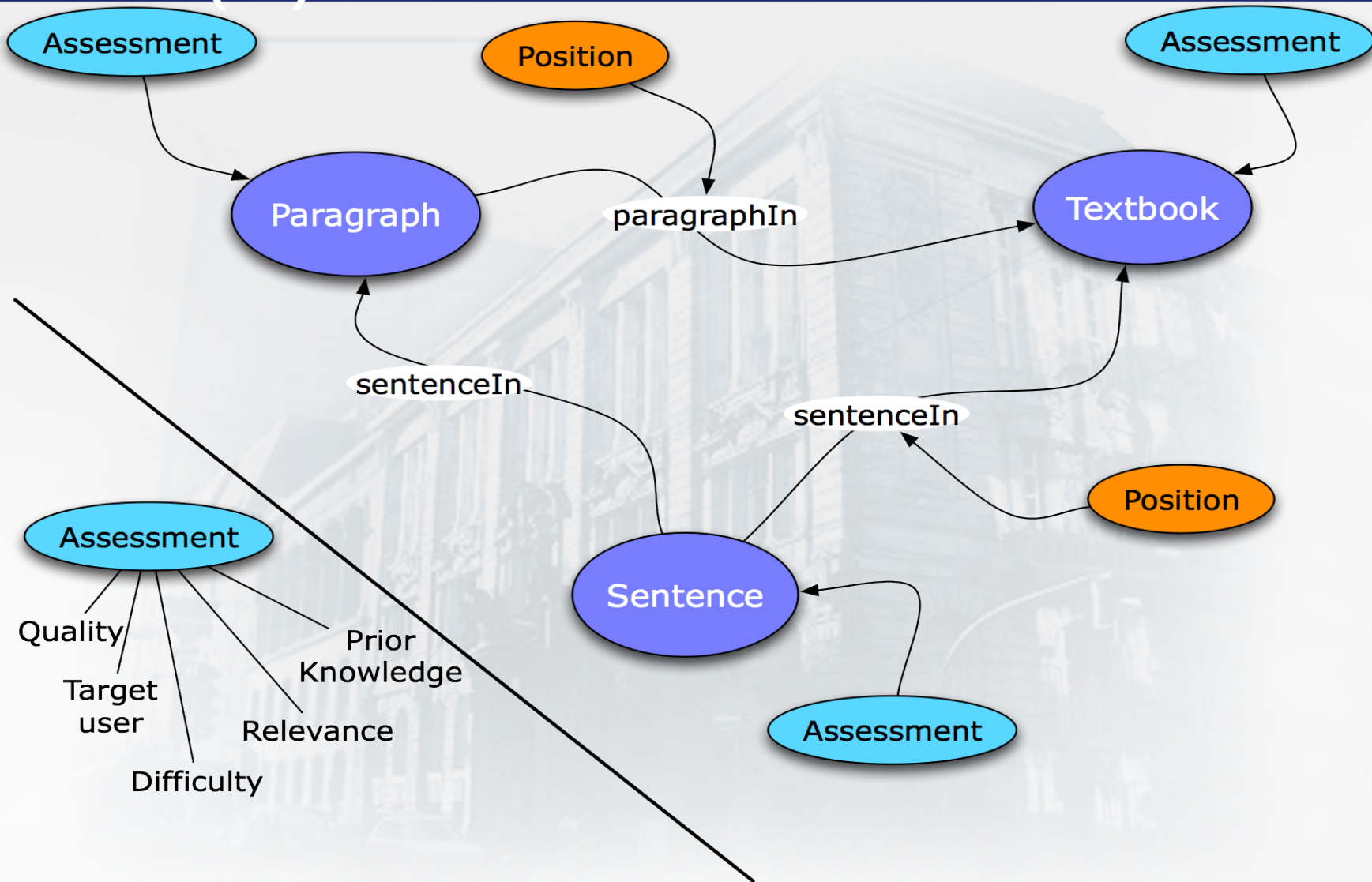
Linked Semantic Publishing: Paux (4)



Linked Semantic Publishing: Paux (5)



Social Semantic Publishing: Paux (6)



Paux live (1): Outline & Sentences



Modul [160345869] PAUX Technologies

Modul Überschriften PAUX-Seiten Tests

Nr.	Eb...	Überschrift	Schlagwörter	Beschreibung	§...	§...	Bst. A...	§...	Bst. A...	G
123	1	English-Abstract	english							
124	2	Introduction (5)								
125	2	History and distribution (3)								
126	2	Comparison with content management systems (8)								
127	3	PAUX-Objects (2)								
128	4	Container objects (contain other PAUX-Objects) (3)								
129	4	Standard objects (1)								
130	3	PAUX-Links (8)								
131	2	Comparison with other Wiki Engines (7)								
132	2	Comparison with other eLearning platforms (9)								
133	2	Comparison with other print-authoring tools (7)								

Details

Medien Überschriften Veranstaltungen Und Hyperlinks Fremd Und PAUX Produkte Bewertung Fundstellen

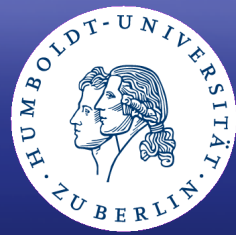
Nr.	Überschrift	Sätze	Lernmöglichkeiten In PAUX	Personen	Fundstellen
1	PAUX is an innovative Content Management System that stores text modularly.				171820175
2	PAUX is software to develop, manage and publicize dynamic individualized content by linking reusable semantic content objects semantically.				91559088
3	PAUX takes a own path of knowledge representation by linking single objects such as words, sentences, pictures, persons etc. to a network.				91559089
4	Therefore, PAUX can be classified only partially in the area of existing systems like CMS/ECMS, Wiki, LMS or LCMS. The semantic content objects of PAUX are to make knowledge available as filterable content for Websites, semantic Wiki, detailed-evaluated eLearning and individualized print media.				91559068
5	PAUX is a knowledge management system written in Java and addresses enterprises (also publishing houses) as well as colleges, offices and other authorities.				91559100
6	It is free of charge in the context of a development partnership.				91559137

NEUER SATZ: Zeile unterhalb anklicken, Strg+N drücken, Speichern bestätigen

Towards Semantic Libraries / Prof. Dr. Stefan Gradmann

Tilburg: Digital Libraries à la Carte 2010

Paux live (2): Sentence & Linking Options



Satz [171820175] PAUX is an innovative Content Management System that stores text modularly.

Satz-Annotationen
Weitere Fundstellen

Historie
Bilder

Durchläufe
Weitere Medien

Infos und Info-Teile
Produkte Und Veranstaltungen

Verknüpfungsübersicht

Aussage
Frage / Antwort
Multiple Choice Test
Lückentext
Lehrbücher
Perso

Satz veröffentlichen

Veröffentlichen ☒ PAUX is an innovative Content Management System that stores text modularly.

Wiederverwendbar ☒

Aussage

Nr. Wort

1	PAUX
2	is
3	an
4	innovative
5	Content Management System
6	that
7	stores
8	text
9	modularly.

Kategorien

Bild	Kategorie
	Sinn & Zweck "Warum-Darum" Didaktik

Grundsatz-Ebene

Anmerkungen

B I

Überschriften

Nr.	Überschrift
124	Introduction (5) PAUX Technologies

Details

Audios
Module
Videos
Dokumente
Produkte
Bilder
Hyperlinks
Animationen

Wort
Sätze
Überschriften
PAUX-Seiten
Tests

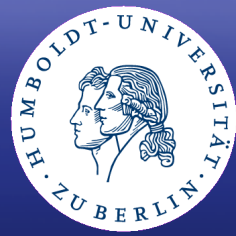
B I

Zeilenwechsel ☐ Fachbegriff ☐ Fragebezug ☐
Wenn ☐ Dann ☐ Weil ☐ So Dass ☐ Um zu ☐
Obwohl ☐ Trotzdem ☐ Vorschriften-Teil ☐

Towards Semantic Libraries / Prof. Dr. Stefan Gradmann

Tilburg: Digital Libraries à la Carte 2010

Paux live (3): Word & Hyperlinks



Satz [171820175] PAUX is an innovative Content Management System that stores text modularly.

Veröffentlichen ☒ PAUX is an innovative Content Management System that stores text modularly.

Wiederverwendbar ☒

Aussage

Nr.	H...	Wort
1	<input type="checkbox"/>	PAUX
2	<input type="checkbox"/>	is
3	<input type="checkbox"/>	an
4	<input checked="" type="checkbox"/>	innovative
5	<input type="checkbox"/>	Content Management System
6	<input type="checkbox"/>	tha
7	<input type="checkbox"/>	stores
8	<input type="checkbox"/>	text
9	<input checked="" type="checkbox"/>	modularly.

Details

Animationen Audios Videos
Dokumente Bilder
PAUX-Seiten Tests Module
Wort Sätze Überschriften
Produkte Hyperlinks

Hyperlink

The Concept of Linked Data (<http://www.w3.org/DesignIssues...>)
Video: "Tim Berners-Lee on the next Web" (<http://www.ted.com...>)

Details

Hyperlink Bewertungen

Hyperlink [The Concept of Linked Data \(http://www.w3.org/DesignIssues...\)](http://www.w3.org/DesignIssues...)

Anmerkung

Kategorien

Bild	Kategorie
	Sinn & Zweck "Warum-Darum" Didaktik

Grundsatz-Ebene

Anmerkungen

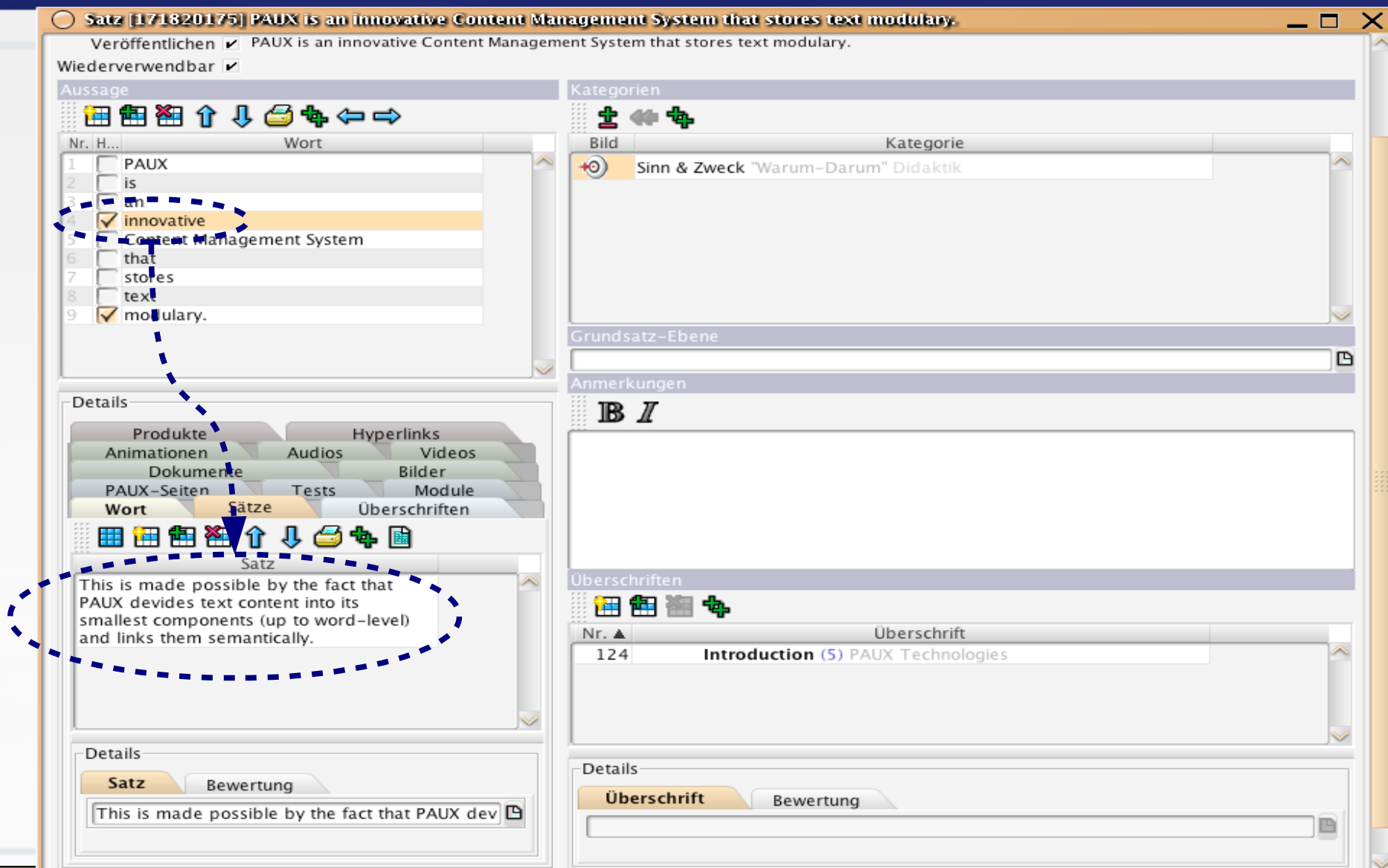
Überschriften

Nr.	Überschrift
124	Introduction (5) PAUX Technologies

Details

Überschrift Bewertung

Paux live (4): Word & Link to Sentence



The screenshot shows the PAUX software interface with the following components:

- Title Bar:** Satz [171820175] PAUX is an innovative Content Management System that stores text modularly.
- Buttons:** Veröffentlichen (checked), Wiederverwendbar (checked).
- Aussage (Statement) Table:**

Nr.	H...	Wort
1	<input type="checkbox"/>	PAUX
2	<input type="checkbox"/>	is
3	<input type="checkbox"/>	an
4	<input checked="" type="checkbox"/>	innovative
5	<input type="checkbox"/>	Content Management System
6	<input type="checkbox"/>	that
7	<input type="checkbox"/>	stores
8	<input type="checkbox"/>	text
9	<input checked="" type="checkbox"/>	modulär.
- Details Panel:**
 - Product categories: Produkte, Animationen, Dokumente, PAUX-Seiten, Wort, Sätze, Überschriften.
 - Hyperlinks: Audios, Videos, Bilder, Tests, Module.
 - Selected item: Satz. Description: "This is made possible by the fact that PAUX devides text content into its smallest components (up to word-level) and links them semantically."
- Kategorien (Categories) Table:**

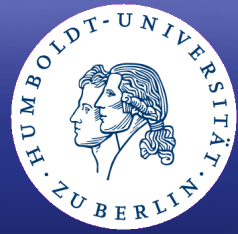
Bild	Kategorie
	Sinn & Zweck "Warum-Darum" Didaktik
- Grundsatz-Ebene (Principle Level):** Empty text field.
- Anmerkungen (Comments):** Text area with bold and italic formatting options.
- Überschriften (Titles) Table:**

Nr.	Überschrift
124	Introduction (5) PAUX Technologies
- Details Panel (Bottom):**
 - Selected item: Überschrift. Description: "This is made possible by the fact that PAUX dev..."

Data = Publication

- Distinction data vs. publication will get **increasingly obsolete** in semantic publishing environments ...
- ... at least in the STM sector.
- The move into semantic publication will be **much slower in the SSH** because of
 - fuzzy and unstable **terminology**
 - fuzzy **linking semantics** hard to formalise consistently
 - close relation between **complex document formats** and **scholarly discourse**
- Current examples are mostly from the medical and bio-medical area as a consequence.
- => Jan Velterop's concept of "Nano-Publications" or Bill Town's examples from chemistry (namely OreChem)

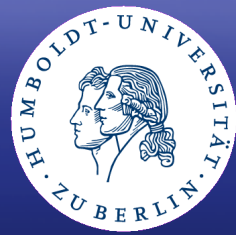
“What do you do with a million books?” (G. Crane)



- DL view: digitisation and (more and more) semantic publishing **increase** by at least one or even more orders of magnitude
 - Scale
 - Linguistic heterogeneity of content
 - Granularity of objects
 - Noise (encoding and semantic)
 - Audience
- They may lead to a dramatic **decrease** of the number of collections and distributors
- They render obsolete the very notion of a '**collection**' ...
- ... as well as the notion of a '**catalogue**'
- → ***Do we need more than one Digital Library in such a setting?***



Re “What do you do with a million books?” (G. Crane)



- Scholarly view: digitisation and (increasingly) semantic publishing result in
 - growing quantity
 - increased complexity
 - Well beyond scholarly processing capacity (=reading faculty)
 - Multiplication of collections or distributors is annoying → as few as possible. Ideally just one (?)
 - Scientists and Scholars will badly need help in these areas:
 - **Semantic abstracting, named entity recognition** for “strategic reading” (Renear)
 - **Contextualisation** of information objects
 - Robust **reasoning and inferencing** yielding digital heuristics
- => ***Opportunities for libraries ... but for others, too!***

A faded, light blue background image of a large, multi-story building with many windows, likely a university building.

Dessert

Who else is going there?

Partners and Contenders

Partners and Contenders: A Mostly Ambivalent Scenario



- Commercial enterprises as **partners**
 - Document mining, named entity recognition and semantic aggregation (OpenCalais, Temis and others)
 - Search engine technology (Google and others)
 - Library automation suppliers (OCLC, ExLibris and others)
- Commercial enterprises as **contenders**
 - Google
 - Amazon
 - Publishers
 - OCLC
- Will other **digital libraries** be partners or contenders?



Conclusion: Towards Semantic Libraries



Another triple paradigm shift

- Focus on **container** → focus on **content**
- Metadata **catalogue** → **semantic network** of contextualized object representations
- Information object **storage** and **retrieval** → **knowledge generation**

Or else the WWW will share the fate of the Concorde aircraft or of the Zeppelin (Eco) and *nothing of this is going to happen* ...

... I don't believe so!



Suggested Reading

- Gregory Crane (2006): What Do you Do with a Million Books? In: Dlib Magazine, Vol. 12, March.
(<http://www.dlib.org/dlib/march06/crane/03crane.html>)
- David Shotton (2009a): Semantic Publishing. The coming revolution in scientific journal publishing. Learned Publishing Volume 22, No 2, 85–94, April 2009; doi:10.1087/2009202
- David Shotton et al. (2009b): Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article
(<http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000361>)
- Barend Mons, Jan Velterop: Nano-Publication in the e-science era
(<http://www.surffoundation.nl/SiteCollectionDocuments/Nano-Publication%20-%20Mons%20-%20Velterop.pdf>)
- Alan Renear, Carol Palmer (2009): Strategic Reading, Ontologies and the Future of scientific Publishing. In: Science, August 2009, p. 828 – 832.

Thank you for your patience and attention